Classifying transaction and web scraped data

Keywords: Scannerdata, web scraping, machine learning, classification

1. INTRODUCTION

A mapping must be established between individual products that are available in scanner/web scraped data sets and the classification used for the statistical product, e.g. COICOP in consumer price statistics. Paper and presentation introduce the range of available (semi-) automated methods (incl. their pros and cons and technical pre-conditions) and describe in detail an advanced approach taken to classify the data for CPI compilation using machine learning techniques.

2. METHODS

"Classification" is one of several working steps when using scanner data and/or webscraping data for CPI compilation purposes. Several international guidelines [1] [2] have been published that describe in general the approaches to be taken when using these new data sources. However, there is a need to develop a better understanding on the different options and methods for the specific working step "classification". The UNs Global Working Group on Big Data for Official Statistics has launched a task team that specifically deals with scanner data cleaning and classification methods and that brings together price statisticians from around the world to develop guidelines and overviews on best practices.

2.1. Overview of the range of methods to classify scanner and webscraped data

The classification of scanner and web scraped data requires NSIs to develop all or most of the following process steps:
-the analysis of retailer specific product groups and item characteristics,
-a manual or automated mapping of most detailed retailer specific product groups to COICOP,
-the exclusion of irrelevant or unusable retailer specific product groups, the identification of relevant item codes within usable product groups,
-the manual and/or automated classification/mapping (if necessary also re-assigning) of relevant item codes to more detailed product classes and sub-classes

The available range of classification methods to manage the process steps range from purely manual to fully automated approaches. Examples of methods & techniques are:
- manual classification (possibly with suggestions based on past classifications)
- (semi) supervised classification
- probabilistic classification
- machine learning methods
- word embeddings

Paper and presentation will introduce and provide an overview on these methods and inform about which NSIs uses/ tests these methods for price index compilation.

2.2. Classification of Web scraped and Scanner Data Using Combinations of Different Machine Learning Techniques

Different Machine Learning (ML) models are tested using scanner data from two major national retailers to find a model that performs with a low error rate (Naive Bayes, Support Vector Machine & Random Forest). First tests show immediately, that there is a practical need to develop an indicator for possible product misclassifications (containing false positives as well as false negatives). Results also show that a certain amount of products still need to be manually checked, but their number should be manageable in terms of work load. By and large, the application of machine learning techniques does not solve all the problems and the development of a sophisticated multi-level classification procedure is necessary.

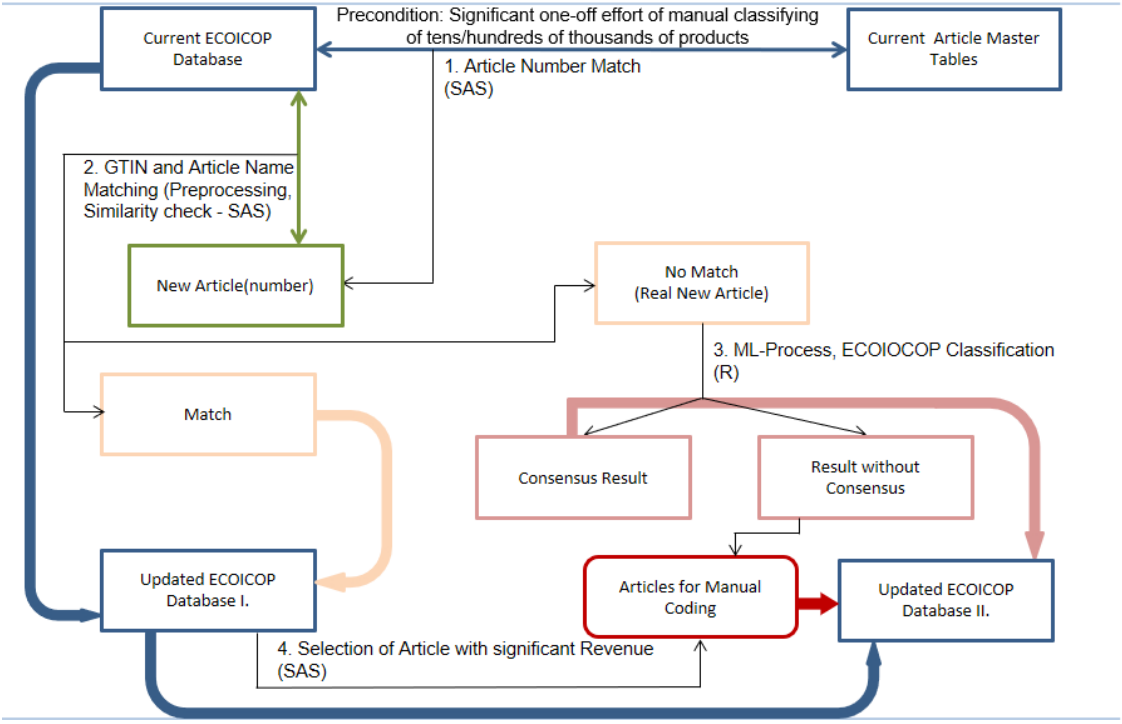

Figure 1. Classification Process of scanner data

3. RESULTS

The consensus between models is overall high. Paper and presentation will describe the applied models and the evolutions of the error rates in different product groups. Findings show that using the three machine learning techniques about 91% of classified data sets are all correct and 4,5% all are wrong. In about 4% of the cases the models delivered diverging results (<- potential of stacking).

CONCLUSIONS

NSIs that built up scanner and web scraped data classification procedures need to take into account their human & technical resources. There is a need to find out if all item codes need to be classified or only the items that are needed for index compilation.

The experiments with Machine Learning methods show that combining the results of more ML models is a feasible way to create a misclassification indicator. In particular, we found that contradicting ML model results can serve as a good indicator of misclassification.

REFERENCES

[1] Eurostat Guide on Scanner data https://circabc.europa.eu/sd/a/8e1333df-ca16-40fc-bc6a-1ce1be37247c/Practical-Guide-Supermarket-Scanner-Data-September-2017.pdf

[2] CPI Manual Chapter 10.15-10.20 https://www.imf.org/~/media/Files/Data/CPI/cpi-manual-concepts-and-methods.ashx?la=en